2021-03

SAFETY ENGINEERING OF WEAPONIZED AUTONOMOUS SYSTEMS

Felder, Javon A.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

CYBER SYSTEMS AND OPERATIONS CAPSTONE REPORT

SAFETY ENGINEERING OF WEAPONIZED AUTONOMOUS SYSTEMS

by

Javon A. Felder

March 2021

Advisor: James B. Michael
Co-Advisor: Loren E. Peitso

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE March 2021	3. REPORT TYPE AND DATES COVERED Cyber Systems and Operations Capstone Report

4. TITLE AND SUBTITLE
SAFETY ENGINEERING OF WEAPONIZED AUTONOMOUS SYSTEMS

5. FUNDING NUMBERS

6. AUTHOR(S) Javon A. Felder

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

This capstone explores the applicability of the Systems-Theoretic Accident Model and Processes (STAMP) framework and the System-Theoretic Process Analysis (STPA) methodology to guide consideration of system safety concerns posed by future variants of Sea Hunter. The author analyzed the Sea Hunter's navigational mission behaviors from a high-level perspective of a functional hierarchy, discussing the specific steps of how basic STAMP/STPA can be used to identify safety hazards and safety hazard casual factors on a complex system such as Sea Hunter. Using the STAMP/STPA methodology, the author provides a functional hierarchy example of the potential system safety hazards involved on the different hierarchy levels in the steering system on Sea Hunter. This capstone discusses how STAMP/STPA can be used to identify system-level hazards, identify unsafe control actions, and identify loss scenarios in the example. The U.S. Navy needs to ensure that its assessment capabilities can be used to adequately identify and evaluate safety hazards, safety hazard causal factors, safety controls, and safety risks of autonomous weapons systems (AWS). AWSs are defined as weapons that can independently select and attack targets. STAMP/STPA is a promising approach to safety analysis; further examination of its applicability and utility in the context of AWS is recommended. If beneficial, this toolset could help the U.S. Navy accelerate the development of fully autonomous technology.

14. SUBJECT TERMS
safety, autonomous weapon systems, STAMP, STPA

15. NUMBER OF PAGES
73

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

SAFETY ENGINEERING OF WEAPONIZED AUTONOMOUS SYSTEMS

LT Javon A. Felder (USN)

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN CYBER SYSTEMS AND OPERATIONS

from the

NAVAL POSTGRADUATE SCHOOL
March 2021

Reviewed by:

James B. Michael
Advisor

Loren E. Peitso
Co-Advisor

Accepted by:

Alex Bordetsky
Chair, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This capstone explores the applicability of the Systems-Theoretic Accident Model and Processes (STAMP) framework and the System-Theoretic Process Analysis (STPA) methodology to guide consideration of system safety concerns posed by future variants of Sea Hunter. The author analyzed the Sea Hunter's navigational mission behaviors from a high-level perspective of a functional hierarchy, discussing the specific steps of how basic STAMP/STPA can be used to identify safety hazards and safety hazard casual factors on a complex system such as Sea Hunter. Using the STAMP/STPA methodology, the author provides a functional hierarchy example of the potential system safety hazards involved on the different hierarchy levels in the steering system on Sea Hunter. This capstone discusses how STAMP/STPA can be used to identify system-level hazards, identify unsafe control actions, and identify loss scenarios in the example. The U.S. Navy needs to ensure that its assessment capabilities can be used to adequately identify and evaluate safety hazards, safety hazard causal factors, safety controls, and safety risks of autonomous weapons systems (AWS). AWSs are defined as weapons that can independently select and attack targets. STAMP/STPA is a promising approach to safety analysis; further examination of its applicability and utility in the context of AWS is recommended. If beneficial, this toolset could help the U.S. Navy accelerate the development of fully autonomous technology.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ACTUV	Anti-Submarine Warfare Continuous Trail Unmanned Vessel
AI	Artificial intelligence
AVIA	Autonomy validation, introspection, and assessment
AWS	Autonomous weapon systems
AWS-L	Lethal autonomous weapon systems
BFTT	Battle Force Tactical Training
BLP	Bell LaPadula security model
CIA	Confidentiality, Integrity, Availability model
CIWS	Close-In Weapons System
COLREGs	International Regulation for Preventing Collisions at Sea
C2	Command and Control
CCW	Convention on Certain Conventional Weapons
DARPA	Defense Advanced Research Projects Agency
DDoS	Distributed Denial-of-Service
DOD	Department of Defense
FMEA	Failure modes and effect analysis
FTA	Fault tree analysis
GAO	Government Accountability Office
GGE	Group of Governmental Experts
HTV	H-II Transfer Vehicle
IHL	International Humanitarian Law
ISS	International Space Station
JAXA	Japan Aerospace Exploration Agency
LOAC	Law of armed conflict
LUSV	Large unmanned surface vehicle
MITM	Man-in-the-middle-attack
MUSV	Medium unmanned surface vehicle
NOSSA	Naval Ordnance Safety and Security Activity
R&D	Research and development
ROE	Rules of engagement

RSCS	Remote Supervisory Control Station
STAMP	System-Theoretic Accident Model and Process
STPA	System-Theoretic Process Analysis
TACON	Tactical control
T&E	Test and evaluation
UAV	Unmanned aerial vehicle
V&V	Verification and validation
VV&A	Verification, validation, and accreditation

ACKNOWLEDGMENTS

I must first thank God above for granting me the fortitude to not only pursue my master's degree from Naval Postgraduate School but to also finish all of my degree requirements. My time here at NPS has been grueling and rewarding in equal measure, and I know God was with me every step of the way.

I am deeply appreciative of the opportunity provided by the U.S. Navy for me to further my education. I thank Professor Michael and Mr. Peitso for their encouragement and wise counsel in helping me organize my thoughts.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

> Machines can do many things, but they cannot create meaning. They cannot answer these questions for us. Machines cannot tell us what we value, what choices we should make. The world we are creating is one that will have intelligent machines in it, but it is not for them. It is a world for us.
>
> —Paul Scharre [1]

Any policy decision to implement fully autonomous weapon systems (AWS) for autonomous platforms will constitute a major paradigm shift in U.S. military operations. Rules of Engagement (ROE) and the Law of Armed Conflict (LOAC) guide U.S. military tactics. Sound decision-making is required when offensively targeting or defensively responding to adversaries' actions. The context is problematic. Even experienced combatants with state-of-the-art decision-support systems face challenges posed by the fog and high tempo of modern warfare. This makes one wonder how well artificially intelligent defense systems will perform the decision-making, acting autonomously or in a cooperative manner, on behalf of a warfighter. If Murphy's Law holds, one can expect something to go wrong, such as mistakenly treating a commercial fishing vessel as a lawful military target.

A commander cannot delegate the legal responsibility for ensuring their forces are faithfully following the LOAC and their command-specific ROE. Under tactical control (TACON), if an employed AWS takes actions that violate the LOAC or command-specific ROE, the commander is still responsible and open to prosecution or disciplinary action [2]. This ultimate responsibility leads commanders to want some level of assurance that an AWS will perform as expected.

The issue of needing to trust in the dependability of semi-automated warfighting systems is nothing new. The U.S. Navy's Aegis Combat System supports engage-on-remote [3], in which the combat system on a ship can be remotely (i.e., not from the host ship) commanded by the Ballistic Missile Defense System (BMDS) to launch against a target [4]. Lack of operational transparency and understanding of semi and fully

autonomous decision-making technology has contributed to a strenuous pushback from commanders at all levels against implementing an automated kill chain [5].

Militaries in different countries are deploying fully autonomous weapons capable of making engagement decisions on their own and controversy surrounds the use of artificial intelligence (AI) used for surveillance and decision-making by weapon systems [5]. The employment of AI in weapon systems raises concerns not just because of perceptions of risk. Such systems are both mission- and safety-critical. From the perspective of system safety, these systems control the release of energy that can harm people, property, and the environment.

If the U.S. military decides to direct the development and field completely autonomous weapon systems, it will need to perform safety-hazards analysis and determine ways to address the hazards and their casual factors in such a way that the systems do not pose unacceptable levels of safety risk. Given that the initial steps of developing full weapon autonomy are underway, it is imperative that system-safety engineering be integrally involved in these developments.

Determining the safety hazards and safety hazard casual factors of fully autonomous weapon systems is a challenging task, such as the lack of transparency of the inner workings of AI-based decision-making. This is known as the AI transparency paradox [6]. No system is perfect, but there needs to be evidence that a system provides the desired level of dependability for particular contexts of use. The body evidence, in the form of a safety case, forms the basis upon which system stakeholders can judge for themselves whether to trust the claims made about the system's safety by the developers and sustainers of the system. This capstone addresses system safety, not mission effectiveness, concerns.

The way AWS are tested and evaluated in the future will play a pivotal role in AWS programs' success. The current and most applicable Department of Defense (DOD) guidance is DOD Directive 3000.9 (Autonomy in Weapon Systems) [7]. This Directive states that systems must go through hardware and software verification and validation (V&V) before being released for operational use. DOD Directive 3000.9 does not discuss

how to assess the growing reliance on AI algorithms and machine learning within an AWS operating environment. The DOD needs to balance the operational effectiveness and system safety of AWS, meaning there will need to be compromises made in order to attain an acceptable level of each.

AWS are disruptive technology [8], the introduction of which will likely significantly alter how the U.S. military operates. Introducing fully autonomous weapon systems into the fleet will impact naval warfighting strategy and tactics. As Gillespie notes, "the changes take several years or decades to have an impact. A technology must be mature with reliable products before it can be trusted in the conflict where many lives depend on its correct performance" [8]. The new capabilities and flexibility that will arise with fully autonomous weapons may need to be introduced incrementally, with the concurrent updating of strategy and tactics. New tactics for an AWS may require the development of new software safety assessment capabilities to ensure AWS can be operated safely in training, maintenance, and operational modes.

The U.S. Navy periodically retools its system safety assessment capabilities in response to technological innovation. A prominent example of this was the transition to the use of software-based no-point, no-fire safety interlocks. Before this transition, the Navy used electromechanical interlocks for weapons mounted on surface ships. The U.S. Navy had to update its safety engineering practices and tools when using software to implement safety requirements [9].

The U.S. Navy is now introducing AI into autonomous vessels. The Navy is a transition partner for the Defense Advanced Research Projects Agency's (DARPA) *Sea Hunter*. The U.S. Navy could one day deploy fully autonomous weapons onto an already fully autonomous ship like the *Sea Hunter*. Best practices for system-safety engineering dictate that safety interests be represented from day one of the development and through the life cycle of a weapon system, regardless of the level of automation of the weapon system [5].

The motivation for conducting the research documented in this capstone is the Navy's keen interest in possible uses of the *Sea Hunter*. The scope of this capstone is

investigating the applicability of the Systems-Theoretic Accident Model and Processes (STAMP) framework [10], the theoretical foundation for the System-Theoretic Process Analysis (STPA) method to guide the consideration of safety concerns posed by fully autonomous naval weapon systems.

In this capstone we address the following questions:

- **Guidance**: What guidance (e.g., instructions, regulations, procedures) already exists with the Department of the Navy for assessing the safety of autonomous systems? What specific guidance addresses the safety of weaponized autonomous systems? Does the Naval Ordnance Safety and Security Activity (NOSSA) have any specific programs to address the assessment of AI-based weaponized autonomous systems? How are foreign navies assessing the safety of their AI-based weaponized autonomous systems?

- **Risks**: What are the potential risks involved in operating naval AI-based weaponized autonomous systems? Are there general classes of safety-relevant requirements for AI-based weaponized autonomous systems? Are existing technical means (e.g., tools, techniques, methods, procedures, processes) for assessing software safety sufficient for assessing the safety of software-intensive AI-based weaponized autonomous systems?

- **Gaps**: Given the answers to questions regarding guidance and risks, what are the capability gaps in assessing the safety of naval AI-based weaponized autonomous systems? What human-capital (e.g., workforce knowledge or experience) gaps exist? What are some ways to resolve those gaps? What are the technical gaps? What are some ways to resolve those gaps?

A. SIGNIFICANCE OF RESEARCH

This research explores ways in which the STAMP/STPA framework can be used to perform safety engineering on AWS, with the aim of informing the revision of existing policy and procedures to address AWS-related system-safety issues.

The U.S. military does not currently have any fully autonomous weapons; the current inventory contains semi-autonomous weapons supervised by human operators to intervene when necessary. A good example of such a system is the U.S. Navy's Phalanx Close-In Weapons System (CIWS). "More than thirty nations already have defensive supervised autonomous weapons for situations in which the speed of engagement is too fast for humans to respond" [1]. CIWS has an automated fire-control system that uses radar data to detect, track, and engage threat objects. A few countries have already developed fully autonomous weapon systems. Israel has developed an anti-radiation drone (IAI Harpy 2) that "can search a wide area for enemy radars and, once it finds one, destroy it without asking permission. It's been sold to a handful of countries, and China has reverse-engineered its own variant" [1]. Other countries like Russia have begun building armed robots and drones for war [5]. With adversary militaries taking the lead in developing fully autonomous weapon systems, the U.S. military will likely respond by developing its own fully autonomous weapon systems—yet another arms race.

The software dependency of fully autonomous weapon systems will create challenges in protection from increasingly sophisticated cyber threats. U.S. military weapon systems are already heavily software dependent and network-enabled, but fully autonomous weapon systems will be completely software dependent.

> Automation and connectivity are fundamental enablers for DOD's modern military capabilities. However, they make weapon systems more vulnerable to cyber-attacks. Although the U.S. Government Accountability Office (GAO) and others have warned of cyber risks for decades, until recently, DOD did not prioritize weapon systems cybersecurity. The DOD is still determining how best to address weapon systems cybersecurity. In operational testing, DOD routinely found mission-critical cyber vulnerabilities in systems that were under development. Yet, program officials GAO met with believed their systems were secure and discounted some test results as unrealistic. [11]

In addition to security, the DOD and U.S. military must create advanced software safety assessment capabilities that will ensure these systems can continue to operate safely when under a cyber-attack; this includes issuing and revising policies and guidance on cybersecurity considerations as they relate to safety on fully autonomous weapon systems.

B. ORGANIZATION AND METHODOLOGY

This capstone uses the DARPA *Sea Hunter* as a case study to determine what needs to be added to current U.S. Navy software safety assessment capabilities to address AI's inclusion into fully autonomous weapon systems. Chapter I discusses the significance of this research. Chapter II discusses current AWS, current U.S. policy on AWS, U.S. plans for complete autonomy, and current international policy on AWS. Chapter III gives an overview of *Sea Hunter's* autonomous capabilities and discusses future variants. Chapter IV discusses generic challenges and vulnerabilities for unmanned medium displacement vessels. Chapter V discusses current software safety capabilities and introduces the STAMP/STPA method. Chapter VI discusses the gaps in current hazard analysis capabilities and suggests that STAMP/STPA could be analyzed for application to *Sea Hunter* and other AWS to mitigate those gaps in current hazard analysis capabilities. Chapter VII contains the conclusions and recommendations for future work.

II. CURRENT AWS PLANS AND POLICY

> Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgement over the use of force.
>
> —Paul Scharre [1]

There is interest in equipping fully autonomous weapon systems with AI technology, as evidenced by the rapid proliferation of open literature on this subject. *Military Applications of Artificial Intelligence* [5] discusses the current and future AI plans for the United States, China, and Russia. As countries begin to adopt and further refine AI applications, it will become apparent how militaries will use AI in autonomous weapons. Current conversation and documentation discuss the legal and ethical concerns of developing fully autonomous weapons. There is little information published on how other countries are currently approaching AI inclusion into fully autonomous weapons.

Military research and development (R&D) planners have an arduous job of making predictions and determining where investments should be made for maximum effect. There are at least three facets to this, but only the first is likely to be fully in the public domain:

1. Technology developments that are happening independently of military spending;
2. Developments which can happen with military investment;
3. Likely developments by potential adversaries [8].

Military R&D planners will be expected to work in tandem with policymakers to ensure future autonomous systems have doctrine in place for successful operations. The United States and other countries are beginning to display high interest in developing fully autonomous weapon systems. The U.S. Navy has demonstrated this with its innovation in developing unmanned underwater vehicles (UUVs) and unmanned surface vessels (USVs), such as DARPA's *Sea Hunter* [12] and Boeing's *Orca* [13]. "Naval analysts believe that it might be possible to acquire hundreds of robotic vessels for the cost of one modern

destroyer. Large capital ships are bound to be prime targets for enemy forces in any future military conflict, while a swarm of robot ships would be more difficult to target and losing even a dozen of them may have a lesser effect on the outcome of combat" [14].

In addition to China and Russia, several more countries are developing and showcasing fully autonomous weapon systems, such as Israel with its *IAI Harpy 2* [15] and South Korea with its *SGR-A1* [16]. Policymakers, in turn, need to identify and weigh the potential for unintended behavior and mishaps. Additionally, ethical and legal concerns are the primary fear of humans' diminishing role in the kill chain. Future policy needs to prioritize the testing phase, explicitly testing the system's reaction to uncommon commands that result in undesirable behavior. The normal routine of military testing to create a system that operates as designed within defined environmental constraints cannot be accepted as a satisfactory end goal with AI inclusion [5].

A. CURRENT U.S. POLICY

In 2012 the DOD its first public policy on autonomy in weapon systems [7]. DOD Directive 3000.09 lays out guidelines for the DOD's development and use of autonomous and semi-autonomous weapon systems. DOD Directive 3000.09 is the first policy document written by any country on fully autonomous weapons. However, the directive does not address all of the potential moral, legal, and operational problems these systems pose. "The directive does not cover autonomous or semi-autonomous cyberspace systems for cyberspace operations; unarmed, unmanned platforms; unguided munitions; munitions manually guided by the operator (e.g., laser or wire-guided munitions); mines; and unexploded ordnance, nor subject them to its guidelines" [17]. The directive requires that all systems be designed to "allow commanders and operators to exercise appropriate levels of human judgment over the use of force" [7]. In section (4/c/3) the directive states: "Autonomous weapon systems may be used to apply non-lethal, non-kinetic force, such as some forms of electronic attack, against material targets" [7]. The non-lethal application of force provides ambiguity. What if non-lethal, non-kinetic force indirectly leads to lethal effects? This type of ambiguity in control of lethal force by fully autonomous systems is

consistent with the fears of many professionals who that domestic and international law should prohibit the development of lethal AWS [5].

Under DOD Directive 3000.09, the U.S. military cannot develop fully autonomous weapon systems unless a waiver is approved. It states that "autonomous or semi-autonomous weapons intended to be used in a lethal manner must be approved by two secretaries of defense and by the Chairman Joint Chiefs of Staff before formal development and again before fielding" [7]. The steep approval chain displays the lack of trust in developing these systems. Under this directive, if fully AWS receives an approved waiver, many of the testing and training requirements may be waived "in cases of urgent military operational need" [7]. The directive does not define the statement "urgent military operational need." There are also many loopholes in the directive, "which may entice decision makers to cut corners on testing and evaluation of AWS in order to realize short-term cost savings or bring the development back on schedule, which could endanger civilians or cause undesirable behavior. The Directive does establish testing requirements that must be complete before the approval of a waiver" [7].

Testing fully AWS with the inclusion of AI will be challenging [11]. Providing realistic conditions and simulating adversary tactics to test an AI system's response to a growing operational environment must be considered a prerequisite. Creating a testing process to meet legal standards will also be challenging, but it remains an end goal. DOD Directive 3000.09 will expire on November 21, 2022, 10 years after it took effect. The U.S. Navy is not currently developing any fully AWS, but the Navy has built the *Sea Hunter*, which may be equipped with fully autonomous weapons someday. Pressure to do so is mounting given that adversaries are already developing and planning to deploy fully autonomous weapon systems. The U.S. Navy needs to upgrade its policy, doctrine, and test-and-evaluation practices to ensure AWS successfully make decisions in place of humans and evolve to handle untested conditions while not violating the laws of armed conflict.

B. U.S. PLANS FOR COMPLETE AUTONOMY

"The Unmanned Systems Integrated Roadmap (FY2011–2036) discusses the visions of all the individual services in pursing technologies and policies that introduce a higher degree of autonomy to reduce the manpower burden" [18]. This plan for autonomy covers developments in ground, air, and underwater systems. The U.S. Navy, Army, Air Force, and Marine Corps all have visions to develop and deploy fully autonomous weapons in the future, yet all lack policy guidelines to ensure dependable and secure operation. Avizienis et al. treat the term "dependability" to include the following attributes: availability, reliability, safety, integrity, and maintainability [19]. Creating additional measures beyond V&V will be an inescapable prerequisite to ensure fully autonomous systems' dependable and secure operation. "To ensure the safety and reliability of autonomous systems and to fully realize the benefits of these systems, new approaches to V&V are required" [18].

The goals for complete autonomy in military systems are ambitious, but it remains to be seen just how valuable the acquisition community and commanders in the field will find them once they are demonstrated [5]. If military applications of AI in the near and distant future progress greatly in target recognition and decision making, complete autonomy will be expected to be utilized as a trustworthy warfighting capability.

C. INTERNATIONAL DISCUSSION

The Convention on Certain Conventional Weapons (CCW) holds annual meetings to discuss legal, ethical, technological, and military standpoints of lethal autonomous weapons systems (AWS-L). In recent years, the Group of Government Experts leading the meetings has not produced any specific AWS-L policy recommendations. The sessions have always concluded with the majority vote that human operators must maintain engagement decisions over AWS-L and that AWS-L are subject to (IHL) [20]. One reason for the lack of progress by the U.N.'s GGE in producing policy guidance is the challenge of coming to agreement on what are the attributes of AWS-L. "Some members of the GGE perceive AWS-L as full autonomy with no manual human control, while other experts view it as still having the option for human control if necessary. The U.S., China, and Russia continue to be the most influential actors and will likely decide if militaries will normalize AWS-L in the future" [20].

III. SEA HUNTER OVERVIEW

> They may carry weapons one day, that's a choice the Navy will have to make, their value is out there and being widely distributed in large numbers, they have to go off by themselves and in harsh, unpredicted environments, they have to sense and make decisions.
>
> —Rear Adm. Nevin Carr [21]

A. CAPABILITIES

The Defense Advanced Research Project Agency (DARPA) initiated the *Sea Hunter* program in 2010 to create an autonomous submarine tracking ship. As the Anti-Submarine Warfare Continuous Trail Unmanned Vessel (ACTUV) progressed over the years, Navy leadership took over the project development as a transition partner. The *Sea Hunter* has a 132 ft length, 16 ft draft, and a displacement of 145 tons and operates continuously, with up to a 10000 mile range, using sonar and other sensors to locate mines and enemy submarines. It has a high-frequency sonar that sends acoustic pings determining the characteristics of potential underwater adversaries to track enemy submarines over a long duration. As technology has evolved in recent years, the Navy changed the *Sea Hunter's* operational scope from teleoperation to increasingly greater levels of autonomy, performing a wide range of functions without requiring human intervention [22]. The Navy integrated *Sea Hunter 1* into a carrier strike group leading a large U.S. Pacific Fleet experiment in the summer of 2020 [23]. Results of the work are unpublished as of the publication date of this report. *Sea Hunter II* will incorporate lessons learned from *Sea Hunter 1* and was scheduled to begin at-sea testing at the end of 2020.

Sea Hunter 1 is the first ship in naval history to set sail and arrive in a distant port without human interaction or oversight. The ship sailed more than 5,200 miles from San Diego, California to Pearl Harbor, Hawaii, and back without a crew's need for navigation and steering. "It demonstrates to the U.S. Navy that autonomy technology is ready to move from the developmental and experimental stages to advanced mission testing" [24]. *Sea Hunter* operates using a Remote Supervisory Control Station (RSCS). The RSCS provides autonomous, independent operations under limited remote human supervision. In Figure 1

[25], the major elements of the *Sea Hunter's* autonomy architecture that support autonomous operation are illustrated. The autonomy architecture consists of a high-level mission planner, health monitor, sensor manager, world model situational awareness component, and intelligent decision support component [25].

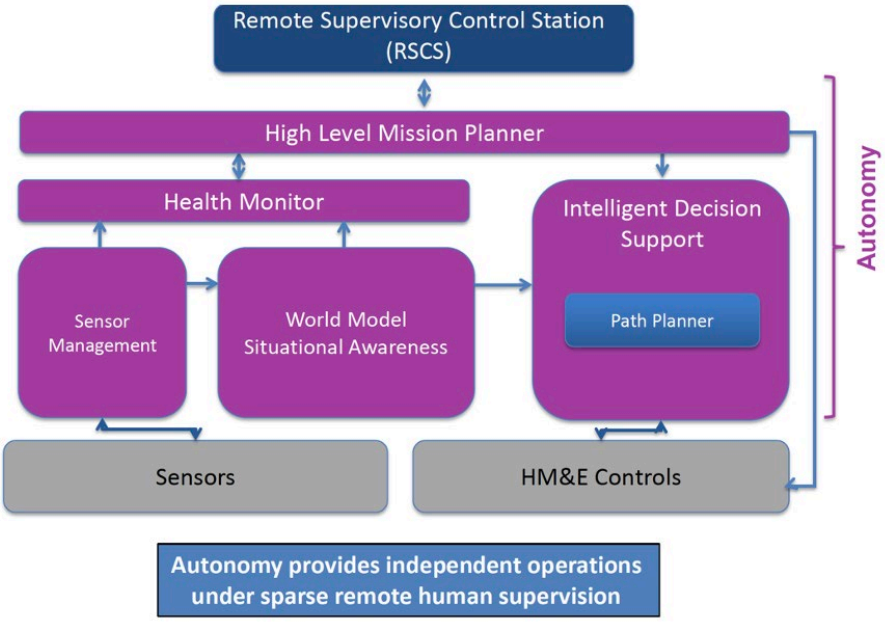


Figure 1. *Sea Hunter A*utonomy Architecture. Source: [25]

These ship's control systems need to perform well under unpredictable environmental conditions for extended periods. Additionally, systems must compensate without external intervention for dependability and security issues that arise, that is, for times when the ship is in a degraded mode of operation.

B. FUTURE VARIANTS

The success displayed by the prototype *Sea Hunter* has resulted in the funding for future variants. According to [26],

> The U.S. Navy awarded L3 Technologies a contract to develop a prototype medium unmanned surface vehicle (MUSV). L3 served as a subcontractor for Leidos, the lead contractor for the Sea Hunter program. L3 plans to

deliver the first MUSV prototype by the end of FY2023. DARPA contracted for two Sea Hunter vessels in what was originally the ACTUV program but has since shifted its focus to be the predecessor to MUSV. A Pentagon office also contracted for two Large USVs as part of the Overlord program, and the Navy will also take those vessels and use them to shape a large unmanned surface vehicle (LUSV) program of its own. The U.S. Navy envisions a family of unmanned systems that will be the backbone of a future fleet of netted "attritable" platforms that will provide lower-cost options compared to manned surface combatants like the Arleigh Burke-class destroyer or the new FFG(X) frigate program. The MUSV and the existing Sea Hunter vessel have different missions and requirements. The existing Sea Hunter vessel was designed and built with the mission of ASW and would be capable of tracking and following submarines using a hull-mounted sonar array over long distances. The MUSV will provide and improve distributed situational awareness in maritime areas of responsibility through intelligence, surveillance, and reconnaissance and electronic warfare implemented by modular payloads.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. GENERIC MUSV CHALLENGES AND VULNERABILITIES

While the following listed challenges and vulnerabilities are not *Sea Hunter* specific examples, *Sea Hunter's* design and implementation must directly deal with these issues. Military Autonomous Weapon Systems are designed to operate within a secure digital perimeter. The resources guarded against outside entities will be strictly "controlled by an outer firewall to prevent unauthorized access" [27]. AWS will complete complex operations and "carry sensitive data, so there is a need to ensure that only authorized" systems and personnel can access these systems [27]. Performing autonomous missions make "it necessary to access and transfer data; the objective of preventing outside individuals from conducting cyber-attacks is achieved by defining and enforcing appropriate access control policies" [27]. Threats to confidentiality in data can be internal or external. By enforcing access control using the Bell LaPadula (BLP) security model, a state machine model used for implementing access control in government and military applications, confidentiality threats by internal threats can be mitigated effectively. Strict implementation of the BLP model "does not allow a user to modify its security label; this ensures that the system never arrives in a state where higher-level information becomes accessible to lower-level users through side channels" [27]. This statement is controversial through as other security experts claim, "covert channels and side channels can occur despite implementation of the BLP model" [28]. Confidentiality of an AWS could also be compromised "by capturing data over network links" [27]. AWS data travels wirelessly and over broad geographical areas, making it simple for cyber-attackers to capture data with sniffers [27]. The data captured with sniffers could consist of "control commands (uplink) and sensor/surveillance data (downlink)" [27]. To date, AWS data that has been compromised was typically a result of weak or non-existent data encryption techniques [27].

Protecting the integrity of AWS from threats depends on the "authenticity of the received data" [27]. Two leading "causes of integrity compromises in AWS could be modified or corrupted data by the hacker before it arrives at the receiver, a man-in-the-middle-attack (MITM)" [27]. The attacker assumes the sender's identity and successfully

sends fake data that appears to be sent from the real sender to the system. Suppose the integrity of a system is not protected. In that case, it can have severe consequences because corrupted data makes incorrect decisions on the system, and the AI mechanisms may not distinguish a corrupt command from intended commands. For example, if a hacker "launches a MITM attack on the Command and Control (C2) data sent to an AWS," it can take control of all functionality [27]. Similarly, if a hacker "modifies the sensor/ surveillance data on a downlink transmission on an AWS supporting a military mission, the effects could be catastrophic" [27]. This is not an abstract fear, Iranian forces took control of and landed a U.S. unmanned aerial vehicle that was conducting surveillance operations near the border [30]. It was reported that some of the communications between the Unmanned Aerial Vehicle (UAV) and ground controllers were unencrypted, which channels were unspecified. This illustrates very directly that an effective, full and comprehensive risk assessment had not been conducted with respect to an active adversary for a real-world deployed UAV system. Similar results are not acceptable for larger and more expensive AWS. What may seem elementary to cybersecurity experts may not be elementary to non-cyber trained systems designers.

An AWS availability can be compromised by hackers gaining unauthorized access and "preventing authorized entities from on-demand and timely access to data and system services" [27]. Future AWS will "utilize autonomously computed commands and cyber services for real-time control of motion, onboard engines, and weapon system suites" [27]. An information warrior's or hacker's objective could be any of the following: to crash software components; "prevent timely delivery to remote commands sent over wireless or satellite links to AWS; or prevent the timely delivery of surveillance/reconnaissance data collected by the AWS" [27]. The leading cybersecurity concerns that can compromise an AWS availability are jamming data communication links, DDoS (Distributed Denial-of-Service) attacks, and injecting malware (viruses, worms, and trojans) into onboard systems [27].

As technology continues to progress, so do our adversaries' strategies and tactics for conducting operations in cyberspace. Analyzing risk within a system in terms of input and output data can also be challenging. Howard and LeBlanc state that "All Input is Evil,"

emphasizing “the point that every input received by a component from the outside can pose a threat to the system” [29] Additionally, in dealing with AWS, any output sent over a wireless transmission poses a threat. The data transmitted wirelessly can be observed by adversaries and is subject to malicious modifications before reaching its destination. Unmanned systems “are being designed with insufficient attention to cybersecurity concerns” [27]. If AWS future is to be secure, we must “provide robust security for protecting data and cyber systems from sophisticated threats” [27].

The emerging fields of AI and Machine Learning (ML) have gained momentum, and the future of military applications in these fields will depend on civilian technology’s success and must be wary of its failures. For example, in February 2019 an innocent man was arrested based on facial recognition software [31]. The software made the wrong decision. Each AWS will operate in complex and dynamic environments and will need to react dependably and securely based on an accurate perception of its surroundings. Each AWS will have multiple sensors employed to make proper decisions with the growing nature of its operating environment. When AWS are fully integrated into the military, it will require that different aspects of systems have a benchmark and proper metrics designated for each component that makes decisions in a human operator’s place. The reference point will set a standard and a baseline for developing the specific systems to help recognize scenarios. For example, if an AWS transits through U.S. waters, it may or may not be scanning for an adversary and it should never mistake civilian crafts or infrastructure for the enemy. This baseline of autonomy will be challenging to maintain with the growing nature of AI and ML.

A kill-chain refers to the sequence of events required to achieve a warfighting effect. In a kill-chain, there are multiple related steps for achieving the desired outcome. The kill-chain does not have a clear doctrinal definition, but it contains a combination of find, fix, and finish across all warfare areas [32].

The prospect of implementing full autonomy into a kill-chain could have a disastrous effect by making the wrong decision, such as noncombatants being injured, killed, or U.S. infrastructure destroyed [32]. This differs from human mistakes, for example, these systems will not have the ability to make ethical decisions, those potential

mistakes create fear among commanders. The decisions these systems make will be limited to the scenarios used to train its sensors [32]. Human operators occasionally make mistakes when making targeting decisions; those mistakes are unacceptable. Regardless of whether a military officer relies on technology to decide and take action, he or she is held accountable. He or she is also responsible for the effects of automated processes, such as those used by autonomous weapons systems. After the Persian Gulf War, Operation Provide Comfort (OPC) was created as a multinational humanitarian effort to relieve the suffering of hundreds of thousands of Kurdish refugees who had fled into the hills of northern Iraq during the war [10]. On patrol supporting OPC, two United States Air Force F-15 Eagle fighters mistakenly shot down two United States Army Black Hawk helicopters that were also participating in OPC and were carrying twenty-six allied personnel [10]. All personnel were killed during the incident. Miscommunication and misunderstanding of the environment were the major factors that led to the improper targeting and shootdown of the helicopters. Human operators still make mistakes while following their command-enforced kill-chain procedures in complex dynamic environments. What guarantees are there that machines are able to operate more accurately as a replacement in those complex dynamic environments? In the future, any lethal-capable AWS will have to execute ROE in parallel with kill-chain responsibilities correctly. Under current kill chains, autonomy is generally employed to replicate items in the kill chain exactly as they are carried out by manned systems [32]. Advances in autonomy have been steady, but the transition to systems capable of reacting to unexpected changes in the environment has not occurred and might not occur for several years [32].

The DOD's "AI strategy directs the DOD to accelerate the adoption of AI and the creation of a force fit for our time" [33]. The future outcomes of war will depend on our ability to use AI to maintain technological and operational superiority over our adversaries. Our pacing competitors are Russia and China, both countries are developing advanced capabilities "such as jamming U.S. military networks and disrupting GPS satellites," [34] and are beginning to achieve these objectives through fully AWS [34]. Simultaneously, the U.S. military has dealt with national and international policy regarding whether weapon systems should operate autonomously and be allowed to use lethal force with no human

oversight. The U.S. military will face several "potential challenges in its future efforts to implement autonomy and AI into military capabilities as an effective deterrent" [34]:

- A near-peer threat gains military edge over the U.S. by being faster to field military capabilities incorporating state-of-the art commercial technology with autonomy or AI.
- The U.S. limits its use of autonomy after interoperability challenges, a particular vulnerability in autonomous weapon systems make autonomous systems less effective than legacy capabilities.
- The U.S. limits its use of lethal autonomy after recurring problems with fratricide, civilian casualties, or other inadvertent engagements.
- The U.S. finds itself lacking freedom of action to use autonomy because:
    - U.S. military operators don't trust and refuse to use autonomous systems, or commanders and political leaders are unwilling to accept the risk.
    - Lethal autonomous weapon systems are preemptively banned by international convention.
    - Our allies refuse to participate in a coalition or provide intelligence to the U.S. if it uses autonomous weapon platforms in military operations [34].

All of these challenges are avoidable, but these concerns have significant consequences if not addressed appropriately. These cut both ways, if sufficient trust cannot be achieved in these systems, they may not be deployed in situations where they can reduce casualties or damage; likewise, if trusted in an unfounded manner, the system may make undesirable decisions which could result in the unintended loss of life or damage to infrastructure.

THIS PAGE INTENTIONALLY LEFT BLANK

V. EXPLORATION OF CURRENT SOFTWARE SAFETY CAPABILITIES

By failing to prepare, you are preparing to fail.

—Benjamin Franklin

The U.S. Navy needs the ability to adequately assess the dependability of their fully autonomously weapons. Otherwise, there is no basis upon which to manage risk or place trust in the dependable operation of these systems. Here we use the terms dependability and trust as defined in [19]:

> The original definition of dependability is the ability to deliver service that can justifiably be trusted. This definition stresses the need for justification of trust. The alternate definition that provides the criterion for deciding if the service is dependable is the **dependability** of a system is the ability to avoid service failures that are more frequent and more severe than is acceptable.
>
> It is usual to say that the dependability of a system should suffice for the dependence being placed on that system. The **dependence** of system A on system B, thus, represents the extent to which system A's dependability is (or would be) affected by that of System B. The concept of dependence leads to that of **trust**, which can very conveniently be defined as *accepted dependence*.

Modern munitions are becoming increasingly networked allowing them to be redirected after initial launch. Traditionally, networked munitions are monitored and retargeted by human oversight. The use of AI will enable the U.S. Navy to transition from relying on human judgement to permit machine intelligence to make decisions about the release and guidance of weapons. "Autonomous systems are complex, and some of the most advanced elements, such as Deep Learning elements, are effectively black boxes" [35]. Their learned algorithms are difficult to uncover, thus making it challenging to perform effective validation, verification, and accreditation (VV&A) [35].

Inadequate VV&A puts Naval personnel in a difficult situation. They need to use the systems they are presented with: They do not have evidence regarding the dependability upon which to objectively decide how much trust to place in the systems. A

commander's ability to trust autonomous systems in their operating environment will be vital to mission success, in addition to mitigating the risks associated with safety and operational hazards. "VV&A is ultimately about mitigating risk. The question is: Can I trust this system to work as planned?" [35] The results of conducting traditional scenario-based testing fails to provide acquisition professionals and operators with sufficient evidence about the dependability attributes of a system. According to Alves [36],

> The context in which autonomous systems will operate makes it hard to even determine the adequacy of test cases. Further, the complexity of the program logic, which is caused by characteristics that are inherent in autonomous systems such as assembling their own course of actions for a given goal, adapting to different environments, learning, diagnosing themselves and reconfiguring themselves to maintain their functionality makes it methodologically and conceptually challenging to assess the coverage of testing on such systems. In fact, a major concern on testing autonomous systems is that the methods are built on the assumption that if a system passes all tests then it will work flawlessly during operation. However, since the behavior of autonomous systems can change over time, it casts doubts on reliability of test results.

Dependability of fully autonomous systems is of paramount importance for many reasons. The systems could be tasked to conduct lengthy missions with only intermittent communications with external command and control systems. Thus, these systems will need to be self-sufficient, performing self-diagnosis of their health and readiness for operation in combat, training, and maintenance modes. In addition to self-diagnosis, any issues encountered will need to be resolved by the autonomous system itself, such as repairs, navigation around obstacles, and even operating safely in degraded modes of operation (e.g., broken bilge pump, inoperable missile hatch, malfunctioning sensors and embedded systems). The bottom line is that any fully autonomous systems to be used by the U.S. Navy will be highly reliant on AI-enabled decision-making in an unpredictably dynamic internal and external environment for which it is impossible to exhaustively test all possible environmental conditions and permutations.

The behavior of the system needs to be such that safety and operational hazards, in conjunction with system failures and environmental conditions, pose an acceptable level of risk. Operational hazards are defined as mishaps or losses, inflicted by outside systems,

particularly hostile forces. System hazards are defined as those conditions that can result in mishaps or losses by actions of the system. "Trust in the autonomy is critical, particularly the software's ability to handle unanticipated failures" [37]. System safety risk needs to be considered in terms of safety hazards, which in turn need to be evaluated in terms of severity of outcome and probability of occurrence. Operational hazards are outside the scope of this capstone, although the information provided in the rest of this chapter is just as applicable for assessing operational hazards as it is for assessing system safety hazards.

A. AUTONOMY VALIDATION, INTROSPECTION, AND ASSESSMENT (AVIA)

AVIA is a modeling and simulation tool developed for rapidly testing and evaluating the logic of fully autonomous systems. AVIA is used to conduct analytic assessments of the behavior of the Sea Hunter's AI system under scenarios in which the state of the operational environment changes. According to [38],

> AVIA's initial objective was to execute a thousand one-hour real-time scenarios in less than 24 hours. One thousand hours of operations is equivalent to 42 days at sea. To reduce 42 days of assessment time to occur in less than 24 hours shows the power of AVIA. This initial objective was achieved in the first year of the program. Subsequent years incorporated improved scenario fidelity, increased metrics for assessing the perception of behavioral logic, and automated approaches to assessing an issue and spinning out additional scenario runs to execute in parallel to the initial 1,000 runs.

AVIA is designed to generate randomized conditions and obstacles, then introduce unexpected events to stress the system. "It can access the actual autonomy logic at several different tap points, so if there's a problem, it can be distinguished from the sensors, the perception, or the behavior response" [38]. AVIA uses the Latin Hypercube Sampling strategy to generate a statistically relevant set of conditions during the testing phase and helps provide confidence that sufficient testing has been conducted to adequately explore the operating space with a near-minimum number of test cases. The safety assessment, along with V&V, needs to focus on the high-severity, high-probability system hazards, and additionally consider the other safety hazards as available resources permit. Identifying

hazards begins with the identification of possible mishaps and working backwards towards potential system attributes, functions, and features that can cause or contribute to mishaps.

AVIA was utilized to test the navigation logic of Sea Hunter and how accurately it followed the International Regulations for Preventing Collisions at Sea (COLREGSs). During Sea Hunter's initial at-sea COLREG testing the vessel exhibited speed and course change indecision. The COLREG testing results were used to make improvements to Sea Hunter's path planner, which "utilizes mission objectives, nautical charts, and sensor data to find the best course to safely reach the vessel's destination" [39]. After the updates to Sea Hunter's path planner, Sea Hunter's ability to accurately follow COLREGs improved significantly. This high proficiency in COLREGS provided the U.S. Navy stakeholders enough evidence for them to place trust in the claims about the dependability of the system, and further, to deploy Sea Hunter from San Diego to Pearl Harbor and back without human intervention [38].

B. SYSTEMS-THEORETIC ACCIDENT MODEL AND PROCESSES (STAMP)

Leveson introduces Systems-Theoretic Accident Model and Processes (STAMP) in *Engineering a Safer World: System Thinking Applied to Safety* [10]. The underlying tenant of STAMP is that "accidents occur when external disturbances, component failures, or dysfunctional interactions among system components are not adequately handled by the control system, that is, they result from inadequate control or enforcement of safety-related constraints on the development, design, and operation of the system" [40]. The goal of STAMP is to prevent future accidents. To achieve this goal, decision-makers and system safety engineers think in terms of "designing and implementing controls that will enforce the adequate safety constraints" [10]. Creating adequate safety constraints and controls that accurately react to those constraints is challenging. STAMP's focus is on "identifying the constraints required to maintain safety; identifying the flaws in the control structure that can lead to an accident (inadequate enforcement of the safety constraints); and then designing a control structure, physical system and operating conditions," all of which enforce the safety constraints [10].

The STAMP causality model is based on system theory and provides the theoretical foundation for System-Theoretic Process Analysis (STPA). A system is defined as a set of system components that act together as a whole to achieve some common goal, objective or end [41]. A system may be part of a larger system or be divided into subsystems. Figure 2 [41], illustrates a typical hierarchy for the system labeled A. The three subsystems (A1, A2, A3) are viewed as systems themselves.

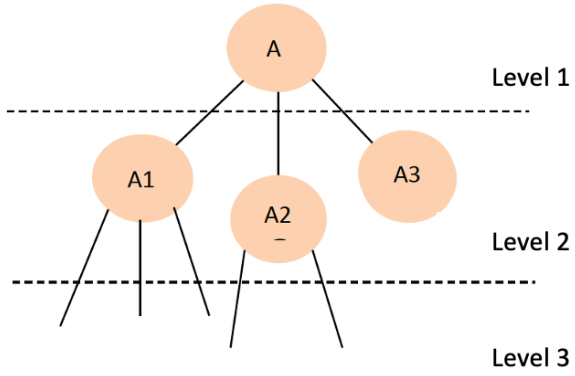


Figure 2. Typical Hierarchy of a System. Source [41].

The advantages of using STAMP are:

- It works on complex systems because it works top-down rather than bottom up [41].
- It includes software, human factors, organizations, safety culture, and casual factors in accidents and other types of losses without having to treat them differently or separately [41].
- It serves as the foundation for tools such as STPA [41].

"In STAMP, systems are viewed as interrelated components kept in a state of dynamic equilibrium by feedback control loops; systems are not treated as static but as

dynamic processes that are continually adapting to achieve their ends and to react to changes in themselves and their environment" [10]. Under the STAMP causality model, when a mishap happens, one or more of the following must have occurred:

1. The safety constraints were not enforced by the controller [10].
    - The control actions necessary to enforce the associated safety constraint at each level of the sociotechnical control structure for the system were not provided.
    - The necessary control actions were provided but at the wrong time (too early or too late) or stopped too soon.
    - Unsafe control actions were provided that caused a violation of the safety constraints.
2. Appropriate control actions were provided but not followed [10].

This causality model is applied differently from one level of the sociotechnical control structure to another. "Each component of the sociotechnical system may have different aspects of an accident under its control and is responsible for different parts of the accident process, that is, different hazards and safety constraints" [10].

C. SYSTEM-THEORETIC PROCESS ANALYSIS (STPA)

Hazard analysis techniques such as Fault Tree Analysis (FTA) and Failure Modes and Effects Analysis (FMEA) [42] were created to analyze primarily electro-mechanical systems and identify potential losses due to component failure. Current software-intensive designs require more powerful analysis approaches that go beyond component failure to identify additional causes of mishaps [5].

STPA is a hazard analysis technique based on STAMP. STPA analyzes component failures and assumes that accidents can also be caused by unsafe interactions of system components, none of which may have failed [41]. Advantages that STPA provides over traditional hazard analysis techniques include:

- Complex systems can be analyzed. “Unknown unknowns” that were previously only found in operations can be identified early in the development process and either eliminated or mitigated [41].

- Unlike the conventional hazard analysis methods, STPA can be used in early concept analysis to assist in identifying safety requirements and constraints, which in turn can then be used to design safety and security into the system architecture and design, eliminating the costly rework involved when flaws in requirements, architecture, or design are identified late in system development or during sustainment and operations. As the system is refined into more detail, the STPA analysis is also refined to inform decisions about detailed design. Complete traceability from requirements to all system artifacts can be easily maintained enhancing system maintainability and evolution [41].

- STPA includes software and human operators in the analysis, ensuring that the hazard analysis includes all potential causal factors in losses [41].

- STPA provides documentation of system functionality that is often missing or difficult to find in large, complex systems [41].

- STPA can be easily integrated into a system engineering process and into model-based system engineering [41].

Comparisons between STPA and traditional hazard analysis methods, such as FTA and FMEA, have been completed on complex systems. In the study titled: Hazard Analysis of Complex Spacecraft Using STPA [42], a comparison between the STPA and traditional methods found that STPA identified all scenarios identified by the traditional methods and also identified additional problematic scenarios [42]. In [42], a hazard analysis of the Japan Aerospace Exploration Agency H-II Transfer Vehicle “KOUNTORI” (HTV) (JAXA HTV) was conducted comparing STPA to traditional hazard analysis methods. The JAXA HTV is an unmanned cargo transfer spacecraft that is launched from the Tanegashima Space Center and delivers supplies to the International Space Station (ISS). This analysis

was conducted by the National Aeronautics and Space Administration (NASA) to identify hazards in approaching, berthing, and departure of the JAXA HTV to the ISS. The results of the analysis stated that "all casual factors identified by FTA were found by STPA, and STPA identified additional casual factors that had not been identified by FTA" [42].

STPA provides a modern hazard analysis technique to address the concerns of new complex technology such as the *Sea Hunter*. Software system safety is still a relatively new safety-engineering discipline, with its body of knowledge and practices evolving rapidly. The software system safety processes being developed are important because software plays a major role in modern weapon systems. For example, software controls of the launching, fuzing, navigation, and guidance of many types of missiles; if there is a design flaw or run-time error, in combination with other physical hazards incorrectly enabling the ability to perform a launch, a mishap might result. Implementing STPA design methodologies into a software-intensive system like the *Sea Hunter* might prove beneficial because no software program is completely error free; nor is it possible to prove that complex software systems are completely error-free. Characterization of software behavior is limited to the operational environment in which it is tested. It is impossible to test every scenario and environment the system software will encounter over its life cycle through inputs from sensors and actuators. Designing with STPA may provide insights into potential hazards early in the design process for future *Sea Hunter* variants. Additionally, STPA has the potential to greatly improve the process of V&V, ensuring the system is designed to the stakeholder's requirements.

VI. DISCUSSION OF GAPS IN CURRENT CAPABILITIES

> There is a widening gap between ambitions and achievements in software engineering. This gap appears in several dimensions: between promises to users and performance achieved by software, between what seems to be ultimately possible and what is achievable now and between estimates of software costs and expenditures.
>
> —Dr. Edward E. David Jr. and Dr. Alexander G. Fraser

A. TESTING AND EVALUATION

To ensure the U.S. Navy keeps its competitive edge over its adversaries, the U.S. is rapidly developing advanced systems, but at the same time ensuring that those systems are dependable and effective. Moving systems and technologies forward through the acquisition process requires conducting sufficient test and evaluation (T&E) to ensure that they can perform their missions in real-world operations [43]. There are concerns about the investment in T&E infrastructure, in July 2017 the U.S. Committee on Appropriations in [43] stated:

> The Committee is concerned that these investments are not being matched by coordinated funding of modern research and testing infrastructure that will ensure that these new systems and technologies are developed and tested as efficiently and quickly as possible and are deployed to operational forces as soon as feasible. The Committee directs the Secretary of Defense to develop a plan for investments in research and testing infrastructure, including through major military construction projects that support development of Third Offset capabilities. The strategy should make clear how the infrastructure investments will be timed so that they are coordinated with planned programs of record.

Third Offset capabilities are defined as U.S. military advantages based on advanced technologies such as AI combined with new operational concepts [43]. Emerging fully autonomous systems will create new challenges for T&E and will require new T&E infrastructure. Fully autonomous systems, like many other systems used with the DOD enterprise, will require regular software upgrades to enhance their capabilities and continuously improve their dependability.

> T&E of autonomous systems needs to involve less-scripted, segmented scenarios to ascertain how well systems perform under specific conditions and rather a broader exploration of how systems perceive and interact with their environments. Autonomous systems will also need to be tested iteratively throughout their life cycles, to reflect software updates and machine learning; a linear set of tests prior to acquisition will not be enough. [43]

The T&E process "will need to be more variable and agile, with the T&E community involved at both earlier and later stages of the life cycle, including system development. The extent of repeatability that is typically sought in T&E of new systems may not be attainable for autonomous systems" [43].

The gap between a developer's ambition and the actual dependability and performance achieved through software is one of the greatest challenges to developing fully autonomous systems because testing a particular set of behaviors does not provide complete trust across all possible behaviors or address potential emergent behaviors. The DOD and military require new T&E infrastructure to conduct extensive testing early in the development stage and periodically throughout the system's life cycle. Identifying and applying the intent of the intellectual property owner is an important step to implementing safe and reliable software into these systems. Additionally, testing the system to safeguard its software from adversary theft by any means is a necessary prerequisite. The DOD and Navy's acceptability criteria for software that shapes the behavior of autonomous systems need to be addressed. The following questions address acceptable criteria for software [8]:

1. Algorithms

- Who owns their intellectual property?
- Does the algorithm have a known mathematical basis?
- Is the mathematical basis known to the weapon system designer?
- What are the assumptions in the mathematical expression of the algorithms?

- Can the algorithms be modelled so that the software implementation can be verified? [8]

2. Software

- Is the software safety or mission critical?
- Is the software sustainable as architected and designed?
- Who owns the intellectual property?
- Will the software or data be highly classified? If so, will there need to be a self-destruct mechanism on any processor or database in a projectile that may fall into hostile hands?
- Has the supplier developed and written the software themselves?
- Is the software under configuration control by the supplier(s)?
- How robust is the software to operating system upgrades?
- What documentation will be supplied with the software?
- Will the customer have access to the code? [8]

An overhaul of military T&E infrastructure would not only require a new vigorous testing process, but it would also require a change in philosophy. A shift in philosophy from identifying software failures during their life cycle to identifying potential software failures during the development stage is necessary for the development of fully autonomous weapon systems. With the end goal for these systems to operate fully autonomously with no human oversight, they must endure a rigorous process of T&E and provide positive results ensuring they are trustworthy to commanders. In addition to creating a T&E process that prioritizes safety over basic acceptable operation, software documentation standards need to be addressed.

Software documentation tends to be inadequate and often times non-existent [8]. Software documentation provides a description of how the system operates and how the

interconnected subsystems are linked together [44]. During maintenance and software upgrade installations, having good software documentation provides developers and users the tools to seamlessly navigate the system [44]. Unfortunately, in most cases, the U.S. military lacks adequate software documentation to assist in troubleshooting. Typically, if software documentation is lacking for a system or difficult to comprehend, the developer is contacted directly, resulting in delayed maintenance and troubleshooting efforts. "To address these issues (at least partially), different approaches and tools have been proposed to aid developers during software documentation, including automatic generation of code and manuals" [44]. From a system-safety engineering perspective, all aspects of development and sustainment need to be documented and tracked, from system conception to retirement and disposal of the system.

Implementing the STAMP/STPA hazard analysis technique into the current T&E problem can enhance the U.S. Navy's ability to assess the dependability of software system safety capabilities. Applying STAMP/STPA to AI-based autonomous systems could provide the Navy's integrated project teams with the information needed about the safety hazards across the entire hierarchy of control systems—information needed to properly assess safety risk and informing the refinement of safety requirements and controls. The next section details this.

B. STAMP/STPA USAGE FOR AWS

As introduced in Chapter 5, the STAMP/STPA process is an approach that could be used to prioritize system safety engineering design for a fully autonomous ship such as the *Sea Hunter*. STAMP/STPA is centered on assessing the system safety engineering based on a functional hierarchy. Figure 3 [45] illustrates a functional hierarchy of high-level operations for the *Sea Hunter*; this hierarchy was created in an analysis of the *Sea Hunter* and its potential contribution to distributed lethality as a Surface Warfare platform [45]. Figure 3 illustrates the functional hierarchy of the potentially lethal *Sea Hunter* design evolution and we can use that as a starting point for determining where to apply STAMP/STPA, thus helping the system safety engineering team explore the impact of the different layered system controls on the safe operation of the *Sea Hunter*. Note that autonomy used

in this sense means that the *Sea Hunter* can perform its tasks without human intervention. It does not mean that there is no human-machine teaming. Specifically, a human operator may remotely take control of *Sea Hunter* if necessary. Further, fully autonomous operation of the *Sea Hunter* may be the exception rather than the norm. The level of autonomy will likely depend on the context of operation and mode of operation (e.g., in combat with degraded or inoperable communications with its tactical controller.

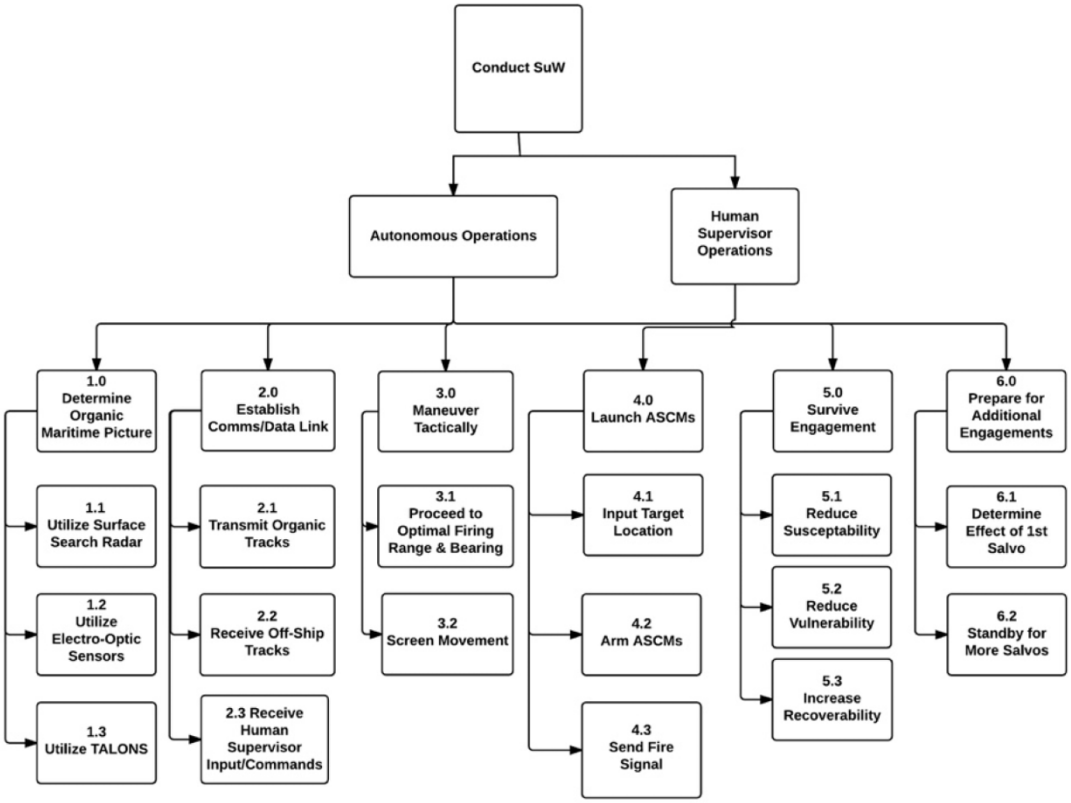


Figure 3. Functional Hierarchy for *Sea Hunter*. Source: [45].

a. Functional hierarchy of high-level potential safety hazards

This section describes a few of the potential safety hazards that can arise with the associated *Sea Hunter* operations described in Figure 3. It is important to distinguish the difference between safety and operational hazards, as they are often confused. As

previously defined, safety hazards are those conditions that can result in mishaps or losses by actions of the system.

- Determine Organic Maritime Picture: The system must accurately identify, track, and understand the difference between friendly and hostile vessels. Misinterpretation of the maritime picture could result in collision or unintended weapons release [45].

- Utilize Surface Search Radar: If the system has a loss of its surface search radar, will it still be able to acquire an accurate maritime picture? Manned vessels can utilize eyesight in the case of this event, but autonomous systems only see through their sensors. Potential safety hazards include collisions with other vessels [45].

- Utilize Electro-Optic Sensors: The loss of electro-optic sensors would be comparable to the loss of the surface search radar because it serves as a backup. If the system is operating in extreme weather, it may not be able to maintain an accurate maritime picture [45].

- Utilize TALONS: The TALON capability extends the range of the surface search radar and the electro-optic sensor by way of an airborne towed payload. Potential safety hazards include the payload crashing into the ocean or potentially the *Sea Hunter* with no human response team available to respond to the casualty [45].

- Establish Comms/Data Link: The *Sea Hunter's* ability to share data with its tactical superior will be crucial if they carry weapons. Under tactical control *Sea Hunter* will likely share track data with nearby ships to increase the tactical picture. Without a steady comms/data link, time-critical firing solutions and weapons release decisions could be lost. This could result in firing a weapon when the vessel should have followed a hold-fire command that arrived or was processed too late [45].

- Receive Human Supervisor Input/Commands: Potential safety hazard casual factors related to receiving human supervisor input/commands include incorrectly processing a command. For example, receiving a command to target an incoming threat, but no response from the system [45].

- Maneuver Tactically: Human operators have the additional advantage of eyesight when maneuvering tactically to engage or evade an adversary. Complete reliance on sensors to maneuver tactically could result in a collision or the inability to maneuver because the system is not comprehending the contact picture or not comprehending it fast enough. Maneuvering tactically happens at high speeds and the radar picture may be slightly delayed. That slight delay can make the difference in accurately avoiding contacts in a congested area [45].

- Proceed to Optimal Firing Range & Bearing: Potential safety hazards that may arise with proceeding to optimal firing range & bearing could be the system fails to produce multiple bearing and range solutions and proceeds to the single solution that is in the same location as a fixed maritime object (hazard of collision) or is in the vicinity of fishing vessels (hazard of injuring civilians) [45].

- Launch ASCMs (Anti-ship cruise missile): The launching of ASCMS may one day be a fully autonomous capability. Potential safety hazards that may arise with launching ASCMs are unintended weapons launch, and mistakenly identifying an ally as an adversary [45].

- Input Target Location: If future ASCM fire control systems are fully autonomous, potential hazards include mistakenly targeting an ally or infrastructure close to the coast, leading to civilian death or fratricide [45].

- Survive the Engagement: Potential safety hazards after surviving an engagement include the accuracy of all damaged components. For

example, if the damaged components are not communicating properly with their subsystems this could lead to collisions, inability to mitigate casualties, or even inadvertent firing of a weapon [45].

- Reduce Susceptibility: The system may not reduce its emissions as intended. For example, if the system is firing ordnance and the radars are not sectored properly it could cause an inadvertent explosion and damage the system [45].

- Increase Recoverability: No manning onboard these systems will be a disadvantage when fighting casualties (not inflicted by an adversary) such as fires and flooding. If the onboard systems are unable to contain the damage, multiple or all systems could be lost. Additionally, this could be an environmental concern with the threat of fuel and oil being discharged into the ocean [45].

In addition to the potential safety hazards that can arise with the associated *Sea Hunter* operations described in Figure 3, safety hazards that can arise during operational training environments and maintenance evolutions are important to mention here. The Battle Force Tactical Training (BFTT) system onboard U.S. Navy warships provides coordinated stimulation of shipboard combat systems to facilitate combat systems team training, providing the capability to conduct realistic joint warfare training across the spectrum of armed conflict and conduct realistic unit-level team training in all primary warfare areas [46]. BFTT accomplishes this by establishing a synthetic environment in which a tactical scenario is run to stimulate shipboard tactical equipment, resulting in coordinated team training events while ships are in-port or underway [46]. Using BFTT, sailors respond to simulated situations by performing multiple types of tasks, including turning keys and press buttons on their combat system equipment as if the ship's crew was in a real-life engagement, building the confidence and gaining the experience needed to successfully engage threats. BFTT can be used by a single ship or be utilized by multiple ships to participate in training evolutions.

Future *Sea Hunter* variants will likely use BFTT to conduct combat system operational training, which in turn may lead to safety hazards arising. For example, BFTT could stimulate the *Sea Hunter* to direct fire on a target, but unless there is a safety interlock which prevents the weapons from releasing ordnance during training, the uncontrolled release of energy could result in a mishap.

Similarly, safety hazards during maintenance evolutions may arise on future *Sea Hunter* variants. For example, on maintenance checks that require human interaction, the system could rotate/radiate weapon mounts or radars, resulting in a mishap on either the unmanned or manned vessels.

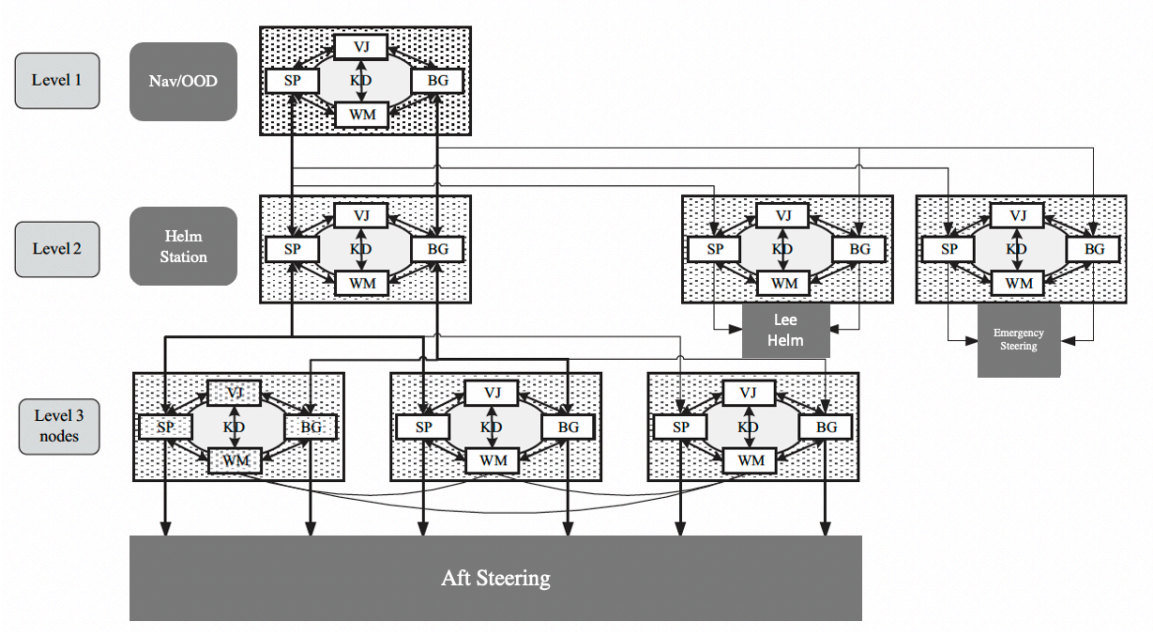
STAMP/STPA might be used to address the safety hazards at multiple levels of the hierarchy as needed for each of the *Sea Hunter's* modes of operation.

b. Multi-level analysis

The previous section discussed potential safety hazards from the high-level perspective of the functional hierarchy. This section will unwrap "3.0 Maneuver Tactically" from Figure 3 and discuss how STAMP/STPA can be applied from highest to the lowest control layer. This analysis can help define the risks and from that the mitigations and or testing requirements. Figure 4 is adapted from [8], it illustrates three hierarchical control layers of steering operations on *Sea Hunter.* Level 1 acts as the top-level supervisory layer making decisions on when rudder movements should occur, think of this as Officer of the Deck and Navigator on a manned ship. Level 2 decides how to implement the commands passed from level 1, think of this as the Helmsman and Lee Helmsman on a manned ship. Level 3 is physically moving the rudder and monitoring the physical components to ensure proper operation, you can compare this to after steering on a manned ship. The illustration in Figure 4 is an implemented representation of the abstract hierarchy from Figure 2, discussed in Chapter IV. The Nav/OOD in Figure 4 is represented as "A" in Figure 2, while the Helm/Lee-Helm and After Steering are the subsystems. The components shown in Figure 4 are the different autonomous sensors that make decisions in level 1, and then pass down necessary instructions to the subsystems to operate all functions of the steering system. The dotted background at each level represents the

network mesh connecting sensors and nodes of the various autonomous sensors. The following describes how the different autonomous sensors would operate in the adapted Figure 4 example [8]:

- The value judgement sensor has pre-set criteria for decisions the steering system will make on rudder adjustments, particularly compensating for environmental effects [8].
- The behavior generator works in tandem with the value judgement sensor and reports all rudder change decisions to the most superior level [8].
- The sensory processing sensor classifies surface targets as military or civilian with high levels of confidence. It will also identify floating debris and other floating objects to safely maneuver around [8].
- The world model sensor has access to relevant information available over communication links. It has the capability to track moving objects in real-time and makes predictions about the contact picture [8].
- The knowledge database provides all knowledge to the system needed to safely navigate as a manned ship would. All decisions and recommendations a Nav/OOD would make will be retained in this database. It is the main sensor that feeds into every decision for the Steering system. The knowledge database at each level functions independently. For example, the knowledge database at level 1 is responsible for making contact picture decisions, while the knowledge database at level 3 simply ensures the physical components move correctly as commanded [8].



[Key: Value Judgement (VJ), Behavior Generator (BG), Sensory Processing (SP), World Model (WM), Knowledge Database (KD)]. Source [8].

Figure 4. *Sea Hunter* Steering Hierarchical Control Layers. Adapted from [8].

Potential safety hazards at each level are:

- Level 1: The system could fail to maneuver in accordance with COLREGS, due to misunderstanding the contact picture or not reacting to it, increasing the chances of collision. The system could fail to recognize that levels 2 and 3 are not executing the ordered commands, resulting in unsafe course corrections due to the delay.

- Level 2: The system could apply the correct degree of rudder but not compensate for weather conditions and sea state, steering the vessel off course. The backup steering system could fail to recognize that the primary steering system is damaged or inoperable.

- Level 3: The system could be sending a signal back to levels 1 and 2 indicating the rudder is moving as intended, but the physical components

have not moved, leaving the rudder fixed and the ship sailing off course, leading to collision or running aground.

These potential safety hazards for the *Sea Hunter* are the same safety hazards experienced on manned ships. The difference is, while the *Sea Hunter* is underway, there are no sailors to contain damage to the vessel or bridge watch standers onboard to navigate by sight if radars and other sensors are inoperable or malfunctioning. Interactions between subsystems may introduce new hazards, for example, communication errors can cause undesired behavior among systems installed on the vessel, or for interaction with other systems within the Fleet in order to perform cooperative-engagement tasks (e.g., launch-on-remote). The implementation of STAMP/STPA to a system like the *Sea Hunter* would inform the evaluation of safety requirements and design and implementation of safety constraints to ensure hazards are less likely to evolve into casualties.

c. STPA Method Overview

As discussed in Ch. 5 STPA can be used on complex systems like *Sea Hunter* to conduct a hazard analysis and identify casual factors of potential accidents. This section will discuss how to conduct a basic STPA analysis as outlined in [41]. The *Sea Hunter* steering system example from Figure 4 will be analyzed in this discussion. The basic STPA method has four steps that are illustrated in Figure 5 [41].

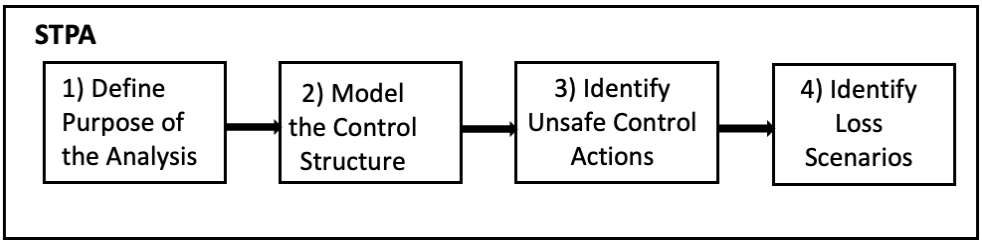


Figure 5. STPA Basic Steps. Adapted from [41].

The first step in the STPA method is to define the purpose of the analysis [41]. The first step is broken down into four parts:

1. Identify losses

2. Identify system-level hazards

3. Identify system-level constraints

4. Refine hazards (optional)

A loss is defined as something of value to stakeholders. Loss includes a loss of human life or human injury, property damage, environmental pollution, loss of mission, loss of reputation, loss or leak of sensitive information, or any other loss that is unacceptable to the stakeholders [41]. For this example, some basic loss scenarios have been discussed in Chapter VI Section B. a. which discusses Figure 3. STPA can be used by stakeholders to prioritize unacceptable losses they want identified during the analysis.

Identifying system-level hazards is the process of identifying the boundaries of that particular system. Specifically, what subsystems are included in the system being analyzed. "The most useful way to define the system boundary for analysis purposes is to include the parts of the system over which the system designers have some control" [41]. For this example, identifying system-level hazards have been discussed in Figure 4.

A system-level constraint is defined as a system condition or behavior that needs to be satisfied to prevent hazards, and ultimately to prevent losses [41]. These constraints can most readily be identified once the system-level hazards are identified. "System-level constraints should not specify a particular solution or implementation. Specifying a particular solution is usually premature at this early stage and can result in alternative and potentially better solutions being overlooked" [41]. For this example, the following system-level constraints have been identified from the hazards discussed in Chapter VI Section B. b. with respect to Figure 4:

- Level 1: The system must maneuver in accordance with COLREGS at all times, this will significantly decrease chances of collision. The system must receive an acknowledgement signal from the subsystems on levels 2 and 3 to ensure rudder commands are executed. This includes signals from

level 2 and/or 3 indicating a degradation to the system that my impact the ability to follow COLREGS.

- Level 2: The system must be able to compensate for incorrect rudder degree due to heavy weather conditions. The backup steering system must be able to recognize when the primary system is damaged or inoperable.

- Level 3: The system must be able verify that the physical components are operating as they are directed to from level 1 and/or level 2.

Refining system-level hazards is the process of detailing the identified hazard into sub-hazards if applicable. Creating sub-hazards can be useful for complex systems such as the *Sea Hunter*. For this example, the following sub-hazards have been identified from level 3 hazards in the Figure 4 discussion:

- Hazard: The system must be able verify that the physical components are operating as they are directed to from level 1. For example, if the command signal is interrupted from level 1 to 3 because of loose components, the physical components will not respond as directed.

    - Sub-hazard: Ensure all steering components are operable jointly and independently, to include the rudder, ram assembly, and lubrication systems. For example, a leak in the lubrication system that could cause gears to seize up.

The second step in the STPA method is to model the hierarchical control structure. For this example, the hierarchical control structure is as illustrated in Figure 4.

The third step in the STPA method is to identify unsafe control actions. An unsafe control action is defined as a control action that, in a particular context and operating environment, will lead to a hazard [41].

Table 1 represents four ways a control action can be unsafe, source [41]:

1. Not providing the control action leads to a hazard.

2. Providing the control action incorrectly in some circumstance(s) leads to a hazard.

3. Providing a potentially safe control action but too early, too late, or in the wrong order.

4. The correct control action is incorrectly applied, e.g., for too long or is stopped too soon.

Table 1. Example of Unsafe Control Actions for the Steering System. Source: [41].

Control Action	**Not providing causes hazard**	**Providing causes hazard**	**Too early, too late, out of order**	**Stopped too soon, applied too long**
Rudder degree	System applies insufficient rudder degree in severe weather conditions to remain on course	System applies correct rudder degree in fair weather conditions System applies insufficient rudder degree in sea state 4 (wind 11–16 knots) or higher	System is late applying sufficient rudder degree, sufficient rudder degree returns in fair weather conditions.	The wrong degree of rudder could be maintained for a long period of time (weather dependent)

Unsafe control actions should specify the context in which the control action is unsafe [41]. For this example, specifying the weather conditions and putting them into context is critical. The weather conditions at sea often require different amounts for the degree of rudder to meet a navigational order. For example, high winds require more rudder to remain on course and favorable currents could require less rudder to remain on course.

For scheduled course changes, this context will be important if the rudder degree is set to a default turning degree for each turn. A small uncompensated offset on a course can put the system in danger of collision with fixed maritime objects or running aground.

The fourth and final step in the STPA method is to identify loss scenarios. A loss scenario is defined as the causal factors that can lead to the unsafe control actions and create hazards [41]. Figure 6 [41] illustrates the relationship between the controller and the controlled process in potential loss scenarios. In this illustration, the controller will act as the level 3 aft steering and the controlled process is the movement of the rudders via the actuators.

1. Why would Unsafe Control Actions occur?

2. Why would control actions be improperly executed or not executed, leading to hazards?

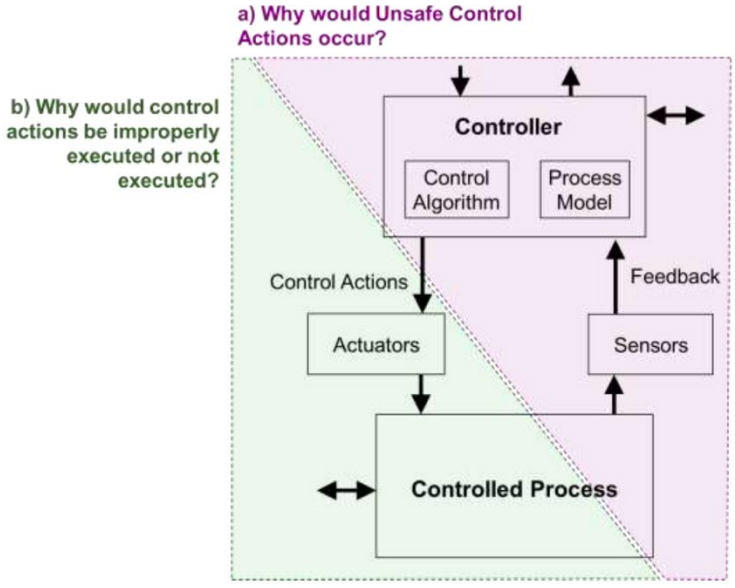


Figure 6. Two Potential Loss Scenarios That Must Be Considered When Assessing STPA Step 4. Source [41].

Loss scenarios that occur due to unsafe control actions may include failures related to the controller, inadequate control algorithm, unsafe control input, and inadequate

process control [41]. For this example, an unsafe control action for the steering system could be because of a power failure. A power failure in the steering system could occur for multiple reasons, including components which have loosened or been displaced out of place due to vibrations or pounding from high speeds or severe weather conditions. A second reason could be lubrication fluids shorting out components, small leaks can cause serious damage over long periods of time. A third reason could be a glitch in the system causing the system to reset and losing power for a short period of time.

Identifying scenarios in which control actions are improperly executed or not executed involves factors that affect the control path as well as factors that affect the controlled process [41]. The control path is defined as the path that transfers control actions to the controlled process. Figure 7 [41], illustrates the control path down to the controlled process. In the current example, the controller in Figure 7 is the aft steering functionality and the controlled process is the movement of the rudders via the actuators. The control path will serve as the downward path from the controller to the controlled process (highlighted in red). Loss scenarios for this example on actions not executed might include a left rudder order from the controller to the actuator, but the actuator does not respond. Loss scenarios for this example on actions improperly executed might include a left 25-degree rudder command, but the rudders only veer left by five degrees.

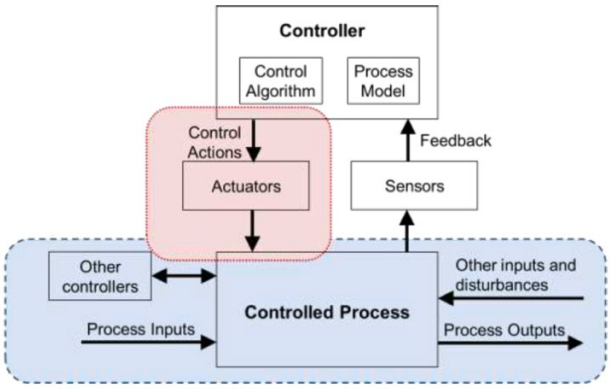


Figure 7. Relationship between Controller and Controlled Process via the Control Path. Source [41].

This level of detail and care is necessary when attempting to prevent safety hazards in fully autonomous weapon systems. Using STAMP/STPA on a system like Sea Hunter, the acquisition team will be able to identify where safety hazards exist in the control hierarchy and identify loss scenarios. Future variants of Sea Hunter will be trusted to operate more safely if a comprehensive and detailed hazard analysis is conducted prior to manufacturing and operational tasking, as well as review and maintenance of the hazard analysis as the system passes through upgrades in its life cycle.

VII. CONCLUSIONS AND FUTURE WORK

> Victorious warriors win first and then go to war, while defeated warriors go to war first and then seek to win
>
> —Sun Tzu, *The Art of War*

As stated in the Introduction, the goal of this capstone was to explore whether the STAMP/STPA methodology might be useful in safety engineering of AWS, with the aim of informing the revision of existing policy and procedures to address AWS related system safety issues. To realize the promise of autonomous technology, the U.S. military must not only aggressively pursue its development, but also create policies that will allow it to be fielded in a timely manner. The U.S. military has several aviation autonomous drones and a fully autonomous Naval ship but does not have the policy and guidance for applying STAMP/STPA or a STAMP/STPA-like methodology to assess the safety of such systems from a sociotechnical, hierarchical control perspective. STAMP/STPA may be a means to bridge the gap, provided that the U.S. Navy includes the framework and methodology—or something akin to it—in its policy and guidance. STAMP/STPA is a possible approach to assessing the dependability of AI-based autonomous systems. This could an organization's ability to provide the evidence that these systems are trustworthy in terms of the safety aspect of dependability.

Some ideas for future research include:

1. Collaborating with NOSSA to implement and test the STAMP/STPA method on a specific weapon system.

2. Collaborating with DARPA and the Navy program office to implement and test the STAMP/STPA method on future variants of the *Sea Hunter.*

3. Collaborating with the DOD to formulate new policy and guidance of software safety engineering, in addition to test and evaluation.

NOSSA, and the acquisition community should conduct further analysis to determine whether applying STAMP/STPA or similar techniques to address software safety in fully autonomous weapon systems would be beneficial.

The *Sea Hunter 1* displayed its ability to navigate and maneuver across the Pacific Ocean with few to no errors. As the *Sea Hunter* ship class begins to increase in numbers, these platforms will become major assets to the Navy fleet.

A question will be asked: Should we develop and equip fully autonomous weapons? We may not have the luxury of a negative answer; the development of fully autonomous weapon systems may occur as result of peer pressure from our adversaries. Using STAMP, future variants of the *Sea Hunter* could be treated as a hierarchical control system, with decisions potentially being made at each level of control. The current policy direction for test and evaluation of autonomous systems focuses on failures, which is a system reliability concern. STAMP or a STAMP-like methodology might support acquisition of dependable—at least in the sense of the system safety—AI-based autonomous systems.

LIST OF REFERENCES

[1] P. Scharre, *Army of None: Autonomous weapons and the Future Of War*. New York, NY, USA: W.W. Norton & Company, 2019.

[2] J. B. Michael, "Trustworthiness of autonomous machines in armed conflict," *IEEE Security & Privacy*, vol. 17, no. 6, pp. 4–6, Nov 2019. [Online]. Available: https://doi.org/10.1109/MSEC.2019.2938195

[3] Assistant Secretary of the Navy for Research, Development & Acquisition, "Cooperative Engagement Capability," Department of Defense. [Online]. Available: https://www.secnav.navy.mil/rda/pages/programs/cec.aspx

[4] C. J. Grant, "CEC: Sensor Netting with Integrated Fire Control," *John Hopkins APL Technical Digest*, vol. 23, no. 2–3, 2002. [Online]. Available: https://www.jhuapl.edu/Content/techdigest/pdf/V23-N2-3/23-02-Grant.pdf

[5] F. E. Morgan, B. Boudreaux, A. J. Lohn, M. Ashby, C. Curriden, K. Klima et al., "Military Applications of Artificial Intelligence," RAND Corp., Santa Monica, CA, USA, 2020. [Online]. Available: https://www.rand.org/pubs/research_reports/RR3139-1.html

[4] C. J. Grant, "CEC: Sensor Netting with Integrated Fire Control," *John Hopkins APL Technical Digest*, vol. 23, no. 2–3, 2002. [Online]. Available: https://www.jhuapl.edu/Content/techdigest/pdf/V23-N2-3/23-02-Grant.pdf

[6] A. Burt, "The AI transparency paradox," Harvard Business Review, December. 13, 2019. [Online]. Available: https://hbr.org/2019/12/the-ai-transparency-paradox

[7] *Autonomy in Weapon Systems*, DOD Directive 3000.09, Department of Defense, Washington, DC, USA, 2012. [Online]. Available: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

[8] T. Gillespie, *Systems Engineering for Ethical Autonomous Systems*. Institution of Engineering and Technology, 2019.

[9] J. Mochulski, R. Malina, "Naval Open Architecture Machinery Control Systems for Next Generation Integrated Power Systems," Washington, DC, USA. 2012. [Online]. Available: https://apps.dtic.mil/dtic/tr/fulltext/u2/a576838.pdf

[10] N. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge, MA, USA: The MIT Press, 2017.

[11] J. M. Inhofe, "Weapons systems cybersecurity*,"* Washington, DC, USA, GAO Report No. GAO-19-128, 2018.

[12] Defense Advanced Research Projects Agency, “ACTUV “Sea Hunter” Prototype transitions to office of naval research for further development,” January 30, 2018. [Online]. Available: https://www.darpa.mil/news-events/2018-01-30a

[13] B. Werner, “Navy awards Boeing $43 million to build for Orca XLUUVs,” USNI News, February 13, 2019. [Online]. Available: https://news.usni.org/2019/02/13/41119

[14] D. B. Larter, “US Navy moves toward unleashing killer robot ships on the world’s oceans,” Defense News, January 15, 2019. [Online]. Available: https://www.defensenews.com/naval/2019/01/15/the-us-navy-moves-toward-unleashing-killer-robot-ships-on-the-worlds-oceans/

[15] IAI, “HARPY,” [Online]. Available: https://www.iai.co.il/p/harpy

[16] NBC News, “Future tech? Autonomous killer robots are already here,” May 14, 2014. [Online]. Available: https://www.nbcnews.com/tech/security/future-tech-autonomous-killer-robots-are-already-here-n105656

[17] K. M. Sayler, “Defense primer: U.S. policy on lethal autonomous weapon systems,” Washington, DC, USA, CRS Report No. IF11150, 2019. [Online]. Available: https://fas.org/sgp/crs/natsec/IF11150.pdf

[18] Department of Defense, “Unmanned systems integrated roadmap FY2011-2036,” Washington, DC, USA, 2011. [Online]. Available: https://fas.org/irp/program/collect/usroadmap2011.pdf

[19] A. Avizienis, J.C. Laprie, B. Randell, and C. Landwehr, “Basic Concepts and Taxonomy of Dependable and Secure Computing,” *IEEE Trans on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11–33, Mar. 2004.

[20] H. Evans and N. Salmanowitz, “Lethal autonomous weapons systems: recent developments,” Lawfare, March 7, 2019. [Online]. Available: https://www.lawfareblog.com/lethal-autonomous-weapons-systems-recent-developments

[21] A. M. Macias, “The Navy offers a glimpse inside the compound where its new drone warship is being built,” CNBC*,* March 11, 2020. [Online]. Available: https://www.cnbc.com/2020/03/11/navy-next-drone-warship-sea-hunter.html

[22] K. Osborn, “Meet the U.S. Navy’s Sea Hunter drone: the robot ship that will hunt submarines,” The National Interest, July 14, 2020, [Online]. Available: https://nationalinterest.org/blog/buzz/meet-us-navys-sea-hunter-drone-robot-ship-will-hunt-submarines-164737

[23] M. Eckstein, “Sea Hunter USV will operate with Carrier Strike Group, as SURFDEVRON plans hefty testing schedule,” USNI News, January 21, 2020. [Online]. Available: news.usni.org/2020/01/21/sea-hunter-usv-will-operate-with-carrier-strike-group-as-surfdevron-plans-hefty-testing-schedule.

[24] J. Trevithick, “Navy’s Sea Hunter Drone Ship has sailed autonomously to Hawaii and back amid talk of new roles,” The Drive, February 4, 2019. [Online]. Available: www.thedrive.com/the-war-zone/26319/usns-sea-hunter-drone-ship-has-sailed-autonomously-to-hawaii-and-back-amid-talk-of-new-roles.

[25] Office of Naval Research, “Medium Displacement Unmanned Surface Vessel,” Washington, DC, USA, 2019.

[26] S. LaGrone, “Navy awards contract for first vessel in is family of unmanned surface vehicles,” USNI News, July 15, 2020. [Online]. Available: news.usni.org/2020/07/14/navy-awards-contract-for-first-vessel-in-its-family-of-unmanned-surface-vehicles.

[27] B. Madan, M. Banik, and D. Bein “Securing unmanned autonomous systems from cyber threats.” Journal of Defense Modeling and Simulation, January 2019. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/1548512916628335?casa_token=lqr-G7bCRQMAAAAA:8LLitafc7Yd_o0CFpqVnFinFydSl-b0Lvt19gSmDAuejZwyqe8igh26FO6-Cz4--ObqHahb4YGmc

[28] C. Irvine, private communication, Nov 2020

[29] M. Howard, D. Leblanc. *Writing secure code.* 2nd ed. Redmond WA: Microsoft Press, 2003.

[30] G. Jaffe, T.Erdbrink, “Iran says it downed U.S. stealth drone; Pentagon acknowledges aircraft downing,” The Washington Post, December 4, 2011. [Online]. Available: https://www.washingtonpost.com/world/national-security/iran-says-it-downed-us-stealth-drone-pentagon-acknowledges-aircraft-downing/2011/12/04/gIQAyxa8TO_story.html

[31] K. Hill, “Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match,” The New York Times, December 29, 2020. [Online]. Available: https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

[32] B. Martin, D. C. Tarraf, T. C. Whitmore et al., “Advancing Autonomous Systems,” RAND Corp., Santa Monica, CA, USA, 2019

[33] Department of Defense, “Summary of the 2018 Department of Defense Artificial Intelligence Strategy,” Washington, DC, USA, 2018. [Online]. Available: https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

[34] L. Lewis, “Insights for the Third Offset: Addressing Challenges of Autonomy and Artificial Intelligence in Military Operations,” Washing, DC, USA, 2017. [Online]. Available: https://apps.dtic.mil/dtic/tr/fulltext/u2/1041043.pdf

[35] Naval Research Advisory Committee, “Autonomous and Unmanned Systems in the Department of the Navy,” Washington, DC, USA, 2017. [Online]. Available: https://www.senedia.org/wp-content/uploads/2018/01/NRAC-Report-Autonomous-and-Unmanned-Systems-in-the-Department-of-Navy.pdf

[36] E. Alves, “Considerations in Assuring Safety of Increasingly Autonomous Systems.” *Assurance Reasoning for Increasingly Autonomous Systems (ARIAS)*, Jul. 2018. [Online]. https://core.ac.uk/download/pdf/161998725.pdf

[37] A. Schultz, “Adaptive testing of autonomous systems,” U.S. Naval Research Laboratory, 2017. [Online]. Available: https://www.nrl.navy.mil/itd/aic/content/adaptive-testing-autonomous-systems.

[38] J. Toon, “AVIA provides test, evaluation for Autonomy Systems,” Georgia Tech Research Institute, September 7, 2016. [Online]. Available: https://www.gtri.gatech.edu/newsroom/avia-provides-test-evaluation-autonomy-systems

[39] Carnegie Mellon University, “NREC contributes to DARPA Autonomous Anti-Submarine Vessel,” National Robotics Engineering Center (NREC), March 25, 2016. [Online]. Available: www.cmu.edu/nrec/media/news-stories-pages/news-stories-2016/darpa-anti-submarine-vessel.html.

[40] N. Leveson, “A New Accident Model for Engineering Safer Systems,” *Massachusetts Institute of Technology,* 2004. [Online]. Available: http://sunnyday.mit.edu/accidents/safetyscience-single.pdf

[41] N. G. Leveson, J. P. Thomas, “STPA Handbook,” March 2018. [Online]. Available:https://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf

[42] I. Takuto, N. G. Leveson, J. P. Thomas et al., “Hazard Analysis of Complex Spacecraft using Systems-Theoretic Process Analysis,” Journal of Spacecraft and Rockets 51, no.2. March 2014. [Online]. Available: http://hdl.handle.net/1721.1/96964

[43] S. Savitz, J. B. Predd, C. S. Adam et al., "Addressing Infrastructure Needs for Test and Evaluation of Autonomous Systems," RAND Corp., Santa Monica, CA, USA, 2020. [Online]. Available: https://cmd.dtic.mil/

[44] E. Aghajani, C. Nagy, O. Lucero Vega-Marquez et al., "Software Documentation Issues Unveiled," *IEEE/ACM 41st ICSE*, 2019. [Online]. doi: https://doi.org/10.1109/ICSE.2019.00122

[45] K. Casola, "System architecture and operational analysis of medium displacement unmanned surface vehicle Sea Hunter as a Surface Warfare component of distributed lethality," M.S. thesis, Dept. of Sys Eng., NPS, Monterey, CA, USA, 2017. [Online]. Available: https://apps.dtic.mil/sti/pdfs/AD1046311.pdf

[46] Naval Sea Systems Command, "AN/USQ-T46 Battle Force Tactical Training (BFTT)," America's Navy, January 15, 2020. [Online]. Available: https://www.navy.mil/Resources/Fact-Files/Display-FactFiles/Article/2166789/anusq-t46-battle-force-tactical-training-bftt/

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California